Journal of Electronic Imaging 12(3), 499–510 (July 2003).

Classification-based macroblock layer rate control for low delay transmission of H.263 video

Ulug Bayazit
Isik University
Electronics Engineering Department
Buyukdere Cad.
34398 Maslak, Istanbul
Turkey
E-mail: bayazit@isikun.edu.tr

Abstract. *Puri and Aravind's method of macroblock bit count estimation for video rate control is based on the classification of the macroblock data into discrete classes and assigning a unique nonlinear estimate for each class and quantization parameter pair. This method stands apart from other methods in the literature, since the model of the bit count versus the quantization parameter relation, parameterized by macroblock variance, is a discrete model generated solely from measurements. We extend their technique for low-delay video rate control (tight buffer regulation) in two ways. We propose a strategy of near-uniform quantization parameter assignments to the macroblocks of a frame that can come close to maximizing an objective spatial quality function, such as PSNR, over the entire frame. We also adaptively update the quantization parameter assignments for the yet to be coded macroblocks, after the encoding of each macroblock, to compensate for any errors in the bit count estimation of the encoded macroblock. Our experiments demonstrate that the proposed rate control method can more accurately control the number of bits expended for a frame, as well as yield a higher objective spatial quality than the method adopted by TMN8.* © 2003 SPIE and IS&T. [DOI: 10.1117/1.1579700]

1 Introduction

For low-delay H.263 video transmission, the fullness of the encoder buffer must be tightly regulated by selecting the appropriate set of encoder parameters to ensure small deviations of the number of actual coding bits from the target number of bits for a small group (we consider one without loss of generality) of video frames. Let us refer to the bits representing the quantized transform coefficients of a macroblock as quantization bits of that macroblock. Let us also refer to the sum of the quantization and overhead bits representing motion vectors and other parameters (coding mode, coding block pattern, quantization parameters) of a macroblock as coding bits of that macroblock. Since the quantization bits account for the majority of the bits in the bitstream, the problem of encoder parameter selection is usually reduced to the problem of the determination of a *quantization parameter* for each macroblock in the frame, which controls the scale of the scalar quantizer applied to the transform coefficients of the macroblock.

If the goal of quantization parameter determination is also the maximization of the quality of the reconstructed frames of the video sequence,[1–3] the macroblock layer rate control problem stated previously is commonly referred to as a bit allocation problem. An accurate optimization requires rate and distortion pairs to be exactly known for all possible quantization parameter assignments to all macroblocks of the frame. Prequantization and precoding of the DCT coefficients of all macroblocks of the frame with all possible quantization parameters to exactly determine these pairs demand high computational complexity, and are therefore not desirable in a low-cost real-time H.263 encoder. The computational complexity of prequantization and precoding could be avoided by employing approximate equations of rate and/or distortion in terms of normalized macroblock energy,[4] quantization parameters,[5–11] or percentage of zeroes among quantized DCT coefficients[12,13] where the unknown parameters in these equations are estimated from empirical data. The equation for the rate is usually used to determine a quantization parameter for the entire frame[4,6,11] or for each macroblock,[5,7] such that the difference between the total bit count estimate and the target bit count for the frame is minimized. A similar goal may also be achieved by estimating from the empirical data the parameters of the generalized Gaussian pdf or the histogram of each DCT coefficient, to select an (average) quantization parameter for a frame, which yields an estimated quantization bin entropy closest to the target bit rate for that frame.[14]

As with most prior art,[4–7,11,14,15] the main goal in this work is the precise approximation of the target number of bits for a frame rather than the maximization of the objective quality of the reconstructed frames. However, this does not mean that the objective frame reconstruction quality should be sacrificed, as demonstrated by the experimental comparisons of the proposed method with the recent macroblock layer rate control method adopted by International Telecommunication Union Telecom Standardization (ITU-T) TMN8,[8–10] which attempts to both maximize the frame reconstruction quality and meet the target bit count

Paper: JEI01045 received Jul. 16, 2001; revised manuscript received Mar. 7, 2002 and Jan. 23, 2003; accepted for publication Jan. 29, 2003.
1017-9909/2003/$15.00

for the frame by employing approximate equations for distortion and rate in terms of the quantization parameter.

Earlier rate control methods of MPEG-2[16] and H.263,[17] which determine the quantization parameters for a macroblock to be encoded based only on the cumulative deviation of the number of actual coding bits from the target bit usage profile, are rather ineffective for tight buffer regulation. Rate control methods adopted by the MPEG-4 verification model,[6,7] ITU-T TMN8,[8–10] and others[4,5,11,14,15] achieve more accurate control due to one or more of the following reasons. 1. As noted before, the relation between the bit count and the quantization parameter is modeled by an equation whose unknown parameters are derived from empirical data and adaptively updated with the actual bit counts measured for the macroblocks. 2. After the encoding of each macroblock, the quantization parameters of the remaining macroblocks are updated to compensate for any errors in the bit count estimation process of the encoded macro-
block. 3. The data content of a macroblock is exploited in the parametric equation.

While the proposed method retains features 2. and 3. as critical ones for tight buffer control, it generates the model of the relation between the quantization parameter and the bit count for a macroblock solely from empirical data, and does not assume any parametric equational form. The parametric rate versus quantization parameter equations used in the literature are either based on certain assumptions about the source[6–10] or on empirical observations of bit counts.[4,5,11] The parametric equation used by He, Kim, and Mitra[12,13] assumes that the bit count depends on the quantization parameter and the macroblock data through a variable representing the percentage of zeroes among the quantized DCT coefficients. The model in the proposed method provides a good fit by avoiding any such assumptions about the source, and also by accommodating substantially higher degrees of freedom in its design process. In the parametric equations that model the bit count versus quantization parameter relation,[4–11] the number of parameters (hence the degrees of design freedom) is limited due to the high complexity required for a design with a large number of parameters. In Chiang and Zhang,[6] the root mean square error (rmse) of the bit count prediction is said to diminish quickly beyond a second-order polynomial approximation to the rate-distortion function. However, the reduction in rmse from a second-order approximation to a third-order approximation is seen to be considerably large (see Chiang and Zhang[6]) for the purpose of tight buffer regulation. Hence, high-order approximations are necessary for good bit count estimation accuracy, even though they may be infeasible. In contrast to the methods that employ parametric equational forms for their models, the model of rate versus quantization parameter relation in the proposed method is comprised of a discrete set of nonlinear minimum mean square error (MMSE) estimates. The design complexity of the model is manageable even when the set is large.

The first stage of the proposed method is the bit count estimation process for each macroblock and quantization parameter pair. A sample statistic representing macroblock activity, such as the standard deviation or the mean of absolute values (of motion compensated difference values) of a macroblock's luminance and chrominance values, is extracted as a feature. Such information is strongly correlated with the number of quantization or coding bits that will be expended when the macroblock is encoded with a given quantization parameter. Macroblock classification is performed by scalar quantization of the extracted feature and by combining the output quantization level with the coding mode. Finally, the combination of macroblock class and macroblock quantization parameter is mapped to a nonlinear estimate of the number of quantization or coding bits that will be expended for the macroblock.

The estimate for a particular combination of class and quantization parameter is designed (trained) by utilizing the knowledge of bit counts measured for previously encoded macroblocks having the same combination. The previously encoded macroblocks could be from the same sequence's past (online training) or different sequences (offline training).

This first stage resembles the method of Puri and Aravind,[15] but we also consider the coding mode for the purpose of macroblock classification. The incorporation of the coding mode information could be critical, since we have observed that conditioning on different coding modes results in nonlinear estimates that are widely different for certain pairs of macroblock classes and quantization parameters.

In the second stage, we employ a conceptually simple strategy of near-uniform assignment of quantization parameters to the macroblocks of a frame to achieve a high objective frame reconstruction quality. This is substantiated by previous work that has experimentally[18] and theoretically[19] demonstrated that the rate-distortion performance of assigning the same quantization parameter to all the macroblocks of a frame is close to that of assigning rate-distortion optimal quantization parameters. Our experiments demonstrate that the objective frame reconstruction quality achieved by this strategy is as good as or even better than that achieved by the strategy of the rate-distortion optimized quantization parameter assignment of TMN8.[8–10]

Near-uniform assignment of the quantization parameters results in the coding of all the macroblocks in a frame at about the same objective quality. In certain applications, it is necessary to define a region of interest (ROI), which needs to be coded at a higher perceptual quality than the background. An elegant way to achieve this goal by means of a visual sensitivity function is presented by Daly, Matthews, and Ribas-Corbera[20] for the determination of quantization parameter weights in a TMN8 encoder. Such an approach yields significant bit rate savings for the same perceptual quality inside the ROI for sequences with significant detail in the background. Perceptual quality optimization inside the ROI was not pursued in our work.

Our work extends Puri and Aravind's classification-based nonlinear estimation method in two major ways. The quantization parameters of the macroblocks of a frame are initialized prior to their encoding by near uniform assignments to achieve high objective reconstruction quality, while the sum of the bit count estimates for the macroblocks very closely approximates the target number of bits for the frame. After the encoding of each macroblock in the frame, the quantization parameters of the remaining mac-

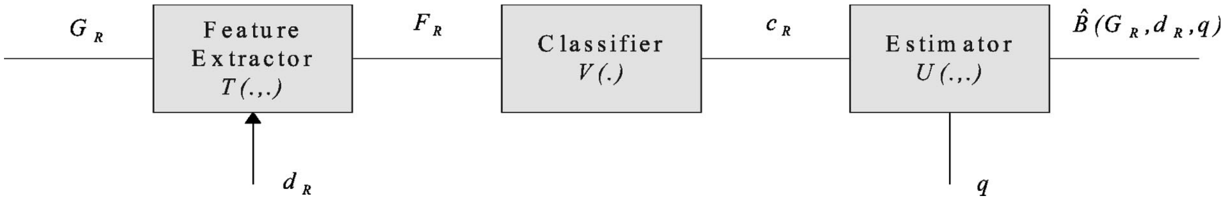

Fig. 1 Block diagram of the bit count estimation process for each macroblock.

roblocks are updated to meet the remaining number of bits for the frame.

The organization of this work is as follows. The bit count estimation process is outlined in general terms as well as by means of a practical example in the next section. Section 2 also discusses the practical design aspects of a nonlinear estimate from empirical data for a particular combination of macroblock class and quantization parameter. In Sec. 3, the strategy of near-uniform quantization parameter assignments to the macroblocks of a frame is described in detail. The methodology of Sec. 2 can be applied to estimate either the number of coding bits or the number of quantization bits and only some of the overhead bits of a macroblock. Section 4 explains these two variations of the proposed rate control method. Experiments providing rate control accuracy and objective frame reconstruction quality comparisons with the rate control methods of TMN5 and TMN8 are presented in Sec. 5. Complexity issues are discussed in Sec. 6.

2 Nonlinear Estimation of the Number of Coding Bits

A functional block diagram for the process of estimation of the number of quantization bits and some or all of the overhead bits to be expended for each macroblock is shown in Fig. 1. While we explain the three major blocks in this estimation process in general terms, we also give a practical example of how an estimate of the number of bits to be expended for an intracoded or an intercoded macroblock can be obtained. The approach outlined here is readily applicable for use in an H.263 or simple profile MPEG-4 encoder, but it can also be adopted for use in other standards-based encoders with trivial modifications.

Let G_R denote the vectorized macroblock luminance and chrominance values (or the motion compensated difference values thereof). Let d_R denote the index assigned to the coding mode of the R'th encoded macroblock in a sequence of coded macroblocks, such that $d_R=0$ when the macroblock is intercoded and $d_R=1$ when it is intracoded. Let q denote the tested value of the quantization parameter for the macroblock. In Fig. 1, the mappings T, V, and U are applied in sequence to yield an estimate $\hat{B}(G_R,d_R,q)$ of the actual number of coding bits $B(G_R,d_R,q)$ for the macroblock. Let the cost of estimating $B(G_R,d_R,q)$ by $\hat{B}(G_R,d_R,q)$ be denoted by $C(B(G_R,d_R,q),\hat{B}(G_R,d_R,q))$. Ideally, these three mappings should be jointly designed to minimize the expected cost

$$\overline{C}(d_R,q)=E[C(B(G_R,d_R,q),\hat{B}(G_R,d_R,q))], \tag{1}$$

for each combination of quantization parameter q and coding mode d_R, where $E[.]$ in Eq. (1) is the expectation of its argument with respect to G_R.

However, in practice, the complexity of such a joint design is prohibitive. In our implementation, the mapping U is the only one designed to minimize the expected cost, while the mappings T and V are determined *a priori*.

At the first step, mapping T extracts the feature vector F_R by acting on G_R and d_R

$$F_R=T(G_R,d_R).$$

Specifically, let us denote the set of pixel locations of the j'th block of the R'th encoded macroblock as $g_{j,R}$. Let $I_{j,R}(x,y)$ be the luminance or chrominance value or the motion compensated difference value thereof at location (x,y) in $g_{j,R}$ with $x\in\{1,\ldots,8\}$, $y\in\{1,\ldots,8\}$, and $j\in\{1,\ldots,6\}$. The mean of the j'th block

$$\overline{I}_{j,R}=\frac{1}{64}\sum_{(x,y)\in g_{j,R}} I_{j,R}(x,y)$$

may be used to compute the rms value of the non-intra-dc coefficients of the R'th macroblock

$$\sigma_R=\left\{\frac{1}{384}\sum_{j\in\{1,\ldots,6\}}\sum_{(x,y)\in g_{j,R}}[I_{j,R}(x,y)-d_R\overline{I}_{j,R}]^2\right\}^{1/2}.$$

In this equation, the means of the blocks of an intercoded macroblock do not have to be subtracted, since the motion-compensated difference macroblocks usually have blocks with very small means.

In our implementation, σ_R and d_R are taken as the components of the feature vector F_R. Such a heuristic choice yields a feature vector F_R of small dimension, which preserves the significant information in G_R correlated with the number of quantization bits to be expended. The fact that the rate-distortion bound for a common source model, such as Gaussian or Laplacian, is parameterized by the source variance is the motivating factor for taking sample statistic σ_R as a feature vector component.

Since the squared error distortion is proportional to q^2, and the distortion is related to rate through the operational rate-distortion characteristics, which we assume is parameterized only by σ_R, the number of bits expended for a macroblock can be expected to be largely determined by the choice for the quantization parameter q. A different sample statistic, such as the mean of absolute values (of motion compensated difference values) of luminance and chrominance values, can also be used if the complexity of computation of σ_R is an issue. On the other hand, the observation that the difference between the number of bits required by intracoded and intercoded macroblocks is large for certain combinations of σ_R and q is the motivating factor for using the coding mode as another feature vector component.

At the second step, classification mapping V maps F_R to a class index c_R

$$c_R = V(F_R), \quad c_R \in \{1, \ldots L\}.$$

In our implementation, this is achieved by quantizing σ_R with a secondary uniform scalar quantizer having bin size δ to get level

$$l_R = \begin{cases} \left\lfloor \dfrac{\sigma_R}{\delta} \right\rfloor & \text{if } \sigma_R < l_{\max}\delta \\ l_{\max} & \text{else} \end{cases}, \tag{2}$$

and combining l_R with mode d_R to get class

$$c_R = V(T(G_R, d_R)) = l_R + d_R(l_{\max}+1)$$

for the R'th encoded macroblock, where $l_{\max}+1$ is the number of levels of the secondary quantizer. The secondary quantizer employed in the classification stage is different than the primary quantizer employed in the main coding loop. Hence macroblocks getting mapped to the same class are of the same coding mode and have similar data content in terms of the sample statistic used.

The classification mapping V applied to the feature vector should result in an insignificant loss of the significant information extracted from the macroblock. In our implementation, δ is small enough such that any further reduction in its value yields insignificant improvement in estimation accuracy (decrease in rms of estimation error is less than 1%). For a given δ, $l_{\max}$ is large enough such that the σ_R for a negligibly small number of macroblocks of the training sequences exceed $l_{\max}\delta$.

At the final step, the mapping U provides an estimate of the number of bits to be expended for the macroblock

$$\hat{B}(G_R, d_R, q) = U(c_R, q).$$

Given T and V, one may rewrite Eq. (1) as

$$\overline{C}(d_R, q) = \sum_{c_R} p(c_R) E[C(B(G_R, d_R, q), U(c_R, q)) | V(T(G_R, d_R)) = c_R], \tag{3}$$

with $p(c_R) = \Pr\{G_R : V(T(G_R, d_R)) = c_R\}$. The design problem of the mapping U is the minimization of $E[C(B(G_R, d_R, q), U(c_R, q)) | V(T(G_R, d_R)) = c_R]$ for any specified q, d_R pair.

Since the rms of the estimation error for the number of bits to be expended is commonly used[7,9] to assess estimation accuracy in rate control, the cost function employed in our implementation is of the form $C(a,b) = (a-b)^2$. For this choice, the cost functional of Eq. (3) is minimized by the class conditional expectation of the number of bits to be expended

$$U(c_R, q) = E[B(G_R, d_R, q) | V(T(G_R, d_R)) = c_R].$$

Let Z be the number of macroblocks in a frame and macroblock R be in the $k+1'$th frame, so that $kZ < R \leqslant (k+1)Z$. By measuring the number of actual bits expended for $P_{kZ}(c_R, q)$ previously encoded macroblocks with indices $(r: 1 \leqslant r \leqslant kZ)$, which are of class $V(T(G_r, d_r)) = c_R$, and are coded with quantization parameter $Q_r = q$, the conditional expectation for $U(c_R, q)$ is estimated as

$$\hat{U}_{kZ}(c_R, q) = \frac{1}{P_{kZ}(c_R, q)} \sum_{\substack{r: 1 \leqslant r \leqslant kZ,\ Q_r = q, \\ V(T(G_r, d_r)) = c_R}} B(G_r, d_r, Q_r), \tag{4}$$

where the updated estimate for mapping $U(\ldots)$ after the encoding of each frame (kZ'th macroblock) is denoted as $\hat{U}_{kZ}(\ldots)$. To refrain from repeating this summation for large k, a recursive update form of the previous equation is used in our implementation

$$\hat{U}_{kZ}(c_R, q) = \frac{\sum_{\left\{\substack{r:(k-1)Z < r \leqslant kZ, \\ V(T(G_r, d_r)) = c_R \\ Q_r = q}\right\}} B(G_r, d_r, Q_r) + P_{(k-1)Z}(c_R, q)\hat{U}_{(k-1)Z}(c_R, q)}{P_{kZ}(c_R, q)}. \tag{5}$$

The summation is over $P_{kZ}(c_R, q) - P_{(k-1)Z}(c_R, q)$ macroblocks in the last encoded frame, which are of class c_R and are coded with quantization parameter q. We have determined that the most recently encoded macroblocks generally yield more accurate estimates. Hence, in Eq. (5), the number of bits expended for the most recently encoded macroblocks are emphasized by applying

$$\left.\begin{array}{l} P_{kZ}(c_R, q) \leftarrow P_{kZ}(c_R, q)/2 \\ P_{(k-1)Z}(c_R, q) \leftarrow P_{(k-1)Z}(c_R, q)/2 \end{array}\right\} \text{ if } P_{kZ}(c_R, q) > P_{\max}. \tag{6}$$

Equation (5) is like a weighted average of the prior estimate of the previous frame and the new estimate based on the observations from the current frame. The weight of the new estimate is

$$1 - \frac{P_{(k-1)Z}(c_R, q)}{P_{kZ}(c_R, q)}$$

and that of the prior estimate is

$$\frac{P_{(k-1)Z}(c_R, q)}{P_{kZ}(c_R, q)}.$$

Note that we allow the weight of the new estimate to be small if the number of macroblocks of class c_R coded with parameter q is small. This way a single macroblock of class c_R coded with parameter q in the new frame cannot excessively degrade the estimate if it is an outlier (bit expenditure does not represent bit expenditures for other macroblocks of class c_R coded with parameter q). On the other hand, we do not allow the weight of the new estimate to be smaller than a constant, $1/P_{\max}$, so that the most recent bit count measurements are prevented from having little effect on the estimate's value.

Especially when $l_{\max}$ is kept large to minimize information loss during classification, it is possible that the training data may not contain any macroblocks of class c_R coded with quantization parameter q $[P_{kZ}(c_R,q)=0]$, and the corresponding estimate $\hat{U}_{kZ}(c_R,q)$ will not be populated. In this case, we approximate $\hat{U}_{kZ}(c_R,q)$ by $\hat{U}_{kZ}(c'_R,q)$ for the class c'_R, which is of the same coding mode ($d'_R = d_R$), closest level (minimizes $|l'_R - l_R|$), and is populated $[P_{kZ}(c'_R,q) \neq 0]$.

3 Quantization Parameter Selection

The technique of bit count estimation presented in the previous section constitutes the basis of the proposed rate control method. Based on this technique, the quantization parameters are determined with low complexity but without rigorous optimization for overall frame reconstruction quality, since such an optimization based on actual rate and distortion pairs is not suitable for real-time encoding implementations with little or no hardware assist, and an optimization based on approximate rate and distortion models[8,9] need not be accurate.

Let us first gain some insight as to why uniform (or near-uniform) quantization parameter assignment may be a good strategy by modeling the bit count versus quantization distortion characteristics of the R'th coded macroblock by the equation

$$R_R(D_R) = \max\left(b, b + a \log \frac{\sigma_R^2}{D_R}\right),$$

where constant b represents the overhead rate and σ_R^2 is the variance of the non-intra-dc coefficients in the macroblock. Considering only coded macroblocks, for which $D_R < \sigma_R^2$, it is straightforward to show that the constant slope operating point condition[21] for optimal rate allocation to macroblocks translates to equal macroblock distortions (i.e., $D_R = D$). This suggests that the quantization parameter, the square of which is proportional to mean-squared error macroblock distortion, should be the same for all the macroblocks of a frame. This result is also supported by Nicoulin *et al.*[19]

In Figs. 2(a) and 2(b), we plot the bit count estimate versus quantization parameter curves for three different classes ($l_R = 0, 10, 50$) obtained by the application of Eq. (4) ($l_{\max} = 100$, $\delta = 4$) on a training set of four sequences coded at various frame rates for intracoded and intercoded macroblocks, respectively. Note that if we exclude the high rate region ($q < 5$) for intracoded macroblocks, the behavior is approximately logarithmic, suggesting that the uniform quantization parameter assignment could be close to optimal in this range.

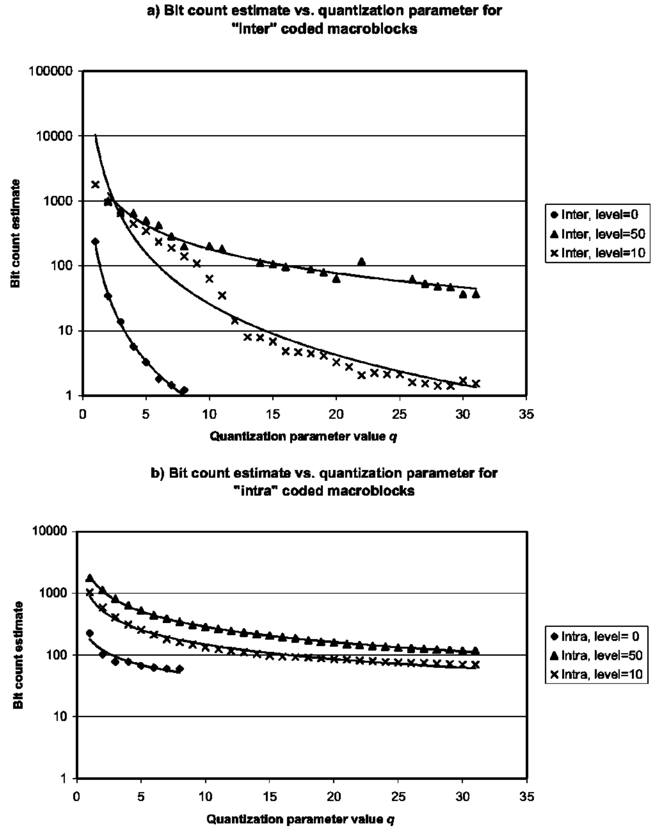

Fig. 2 Plots of the bit count estimate versus quantization parameter value. Each curve is a best polynomial fit for data points of a different macroblock class (indicated on the right). The curves are approximately linear for quantization parameter values larger than 5, suggesting that the logarithmic modeling of rate-distortion function is valid, and uniform assignments of quantization parameters to macroblocks should be close to optimal in this range.

Initially, in our implementation, value q_1 is assigned to the quantization parameters of the first Z_0 macroblocks of a frame (in a predetermined scan order) and value $q_2 = q_1 + 1$ is assigned to the quantization parameters of the remaining $Z - Z_0$ macroblocks to achieve near-uniform spatial reconstruction quality. Raster scan order and reverse raster scan order are alternately employed from frame to frame to achieve subjective temporal smoothness. To determine the best q_1, Z_0 pair, q_1^*, Z_0^*, we first set $q_1 = 30$, $Z_0 = 0$, and then iteratively increment Z_0 (modulo Z) and decrement q_1 each time $Z_0 = 0$ until the difference between the target number of coding bits and the estimated number of coding bits for the macroblocks of the frame is minimized.

After the encoding of each macroblock, we update the values for the quantization parameters in a similar fashion until the difference between the target number of coding bits and the estimated number of coding bits for the *remaining* macroblocks of the frame is minimized.

Specifically, let Z_1 denote the number of encoded macroblocks in the $k+1$'th encoded frame, and B_{TR} denote the target number of bits for the remaining $Z - Z_1$ macroblocks.

The best pair for the raster scan order is given by

$$q_1^*,Z_0^* = \underset{\{q_1,Z_0: q_1\in(1,\ldots,30),Z_0\in(0,\ldots,Z-1)\}}{\operatorname{argmin}} \left| \sum_{\substack{r\in\{kZ+Z_1+1,\ldots,\\ kZ+Z_0\}}} U_r(c_r,q_1) + \sum_{\substack{r\in\{kZ+Z_0+1,\ldots,\\ (k+1)Z\}}} U_r(c_r,q_2) - B_{TR} \right|. \quad (7)$$

A similar expression may be used to determine the best pair for the reverse raster scan order.

Each time a macroblock is encoded, B_{TR}, the available bit budget, is reduced by the number of bits expended to encode that macroblock, and Z_1 is incremented by one. Reapplication of Eq. (7) (for raster scan order) adjusts q_1^* and Z_0^* ($Z_0^* > Z_1$), so that the total number of bit count estimates for the remaining macroblocks [macroblocks with indices $kZ+Z_1+1$ to $(k+1)Z$] in the $k+1$'th frame is as close as possible to B_{TR}. This adaptive updating of the quantization parameters prevents bit count estimation errors from accumulating.

4 Estimation of the Number of Quantization Bits

The rate control method introduced in the previous sections can be implemented by estimating the number of quantization bits and all or some of the overhead bits expected to be expended for each macroblock. The estimate for class c and quantization parameter q can be trained with the sum of all quantization bits and all or some overhead bits expended for macroblocks of class c and coded with quantization parameter q. When the sum of all overhead bits and all quantization bits is estimated, we call this basic algorithm "***New***." In another variation of the algorithm, the sum of only some overhead bits and all the quantization bits is estimated, and this estimate is combined with the *measured* number of remaining overhead bits to yield the number of coding bits expected to be expended. We call this algorithm ***New***$^\kappa$.

In our software encoder implementation, which is similar to the baseline (annex modes turned off) TMN5 software encoder provided by Telenor Research, Norway, the number of overhead bits for H.263 data fields COD, MCBPC, CBPY, and DQUANT, which depend on the quantization levels and are not available prior to the selection of the quantization parameters, have to be estimated. The number of bits for the motion vectors can be measured prior to the selection of the quantization parameters and do not have to be estimated. However, the measurement of the number of overhead bits for the motion vectors requires the differential motion vectors to be formed and coded. This might increase storage and computational complexities partly because the bits representing differential motion vectors cannot be readily placed onto the bitstream. The data partitioning error recovery option of MPEG-4 video, which places the bits for all the motion vectors before the bits for the quantized transform coefficients in the bitstream, allows the total number of motion vector bits to be measured before the quantization parameters are determined, and eliminates the need for additional manipulation and storage of motion vector bits. In other possible encoder implementations, the number of bits for other overhead data elements can be measured beforehand, if the dependencies of these elements on the quantization parameters are removed.

Let us consider the cost for each class and quantization parameter for the direct estimation of the number of coding bits (the sum of all quantization bits and all overhead bits)

$$\begin{aligned}\bar{C}(c,q) &= E[C(B(G_R,d_R,q),U(c_R,q))|V(T(G_R,d_R))=c_R],\\ &= E[(B(G_R,d_R,q)-U(c_R,q))^2|V(T(G_R,d_R))=c_R],\\ &= E[(B-E[B|c_R])^2|c_R],\end{aligned}$$

where we have substituted the optimal estimate for $U(c_R,q)$ for the squared error cost function and also have used the shorthand $B \triangleq B(G_R,d_R,q)$. Similarly, the cost for each class and quantization parameter for the estimation of the sum of all quantization bits and some overhead bits may be expressed as

$$\bar{C}^\kappa(c_R,q) = E[(B^\kappa - E[B^\kappa|c_R])^2|c_R],$$

where $B^\kappa = B - B^\gamma$ denotes the sum of all quantization bits and some overhead bits for a macroblock of class c_R and quantization parameter q. Since the number of remaining overhead bits B^γ can be directly measured, $\bar{C}^\kappa(c_R,q)$ is also the total cost for this case. The difference between the cost expressions can be shown to be

$$\bar{C}(c_R,q) - \bar{C}^\kappa(c_R,q) = 2\rho_{B^\kappa B^\gamma}\sigma_{B^\kappa}\sigma_{B^\gamma} + \sigma_{B^\gamma}^2,$$

where $\rho_{B^\kappa B^\gamma}$ is the correlation coefficient between the random variables B^κ and B^γ. The advantage of estimating B^κ is best realized if the correlation coefficient is 1, and there will clearly be an advantage if the correlation coefficient is nonnegative. Nonnegativity of the correlation coefficient is seen to be generally true if we consider that within an image sequence, fast-motion spatial regions with high motion vector rates yield a higher expenditure of quantization bits when coded for the same reconstruction quality as the slow motion or still regions with low-motion vector rates. It is not unreasonable to assume that motion-compensated prediction works at its best for slow or no motion.

If B^κ is estimated, then the target number of bits B_{TR}^κ is used in place of B_{TR} in Eq. (7), where

$$B_{TR}^\kappa = B_{TR} - \sum_{r\in\{kZ+Z_1+1,\ldots,(k+1)Z\}} B_r^\gamma.$$

We refer to the variation of the algorithm in which an estimate for B^κ is added to a direct measurement for B^γ to yield an estimate for B as ***New***$^\kappa$ to distinguish it from the basic algorithm ***New***, in which an estimate for B is directly obtained.

5 Experiments and Results

A series of experiments was performed with a baseline H.263 TMN5 software encoder compiled with a GNU C Compiler V. 2.95.3 on a SuSE Linux 7.2 workstation equipped with a Pentium III 733-Mhz CPU and 256-Mbytes RAM. Seven QCIF format test image sequences Akiyo, Carphone, Claire, Coastguard, Container, Foreman,

Table 1 Root mean square of the deviation of the number of actual bits from the target number of bits for a frame.

	Kbits/sec, Hz	Akiyo	Carphone	Claire	Coast	Container	Foreman	Silent
TMN5	48, 10	522.9074	447.9988	508.2179	322.0074	843.6012	599.6318	570.2511
TMN8	48, 10	67.30187	67.38460	31.35741	25.09723	153.5764	154.2529	50.18723
New	48, 10	51.38614	38.54302	32.38018	73.13922	57.98741	139.2940	40.80232
New$^{\kappa}$	48, 10	63.91765	21.34540	34.59150	27.76270	44.18353	126.2758	36.03157
New (No update)	48,10	769.6659	555.6503	573.4841	451.9917	417.7489	674.2913	615.0949
TMN5	128, 30	766.4395	687.7349	756.6903	528.9618	820.5396	477.6827	631.2123
TMN8	128, 30	102.4280	30.32600	89.10433	23.47954	263.0058	62.41379	44.03354
New	128, 30	58.08786	21.39399	45.47047	41.6393	51.06416	41.54713	38.94909
New$^{\kappa}$	128, 30	63.63928	20.34995	41.29809	17.27734	53.35861	48.30791	43.29882
New (No update)	128,30	675.3783	376.2962	480.9339	344.6069	344.5858	433.2839	538.0238

and Silent were coded at 48 Kbits/sec and 10 frames/sec, and 128 Kbits/sec and 30 frames/sec. All coded sequences were 10 s long except for Silent, which was 15 s long. The first frame of each sequence was coded as an I frame (with quantization parameter set to 15), and the remaining frames were coded as P frames. The frame layer rate control method of TMN8[9] was employed.

In the TMN8 frame layer rate control method, the encoder buffer is emptied out at a rate of R/F bits per frame interval, where R and F are the channel and frame rates, respectively. For as long as the number of bits in the encoder buffer W is larger than a maximum value M after emptying out R/F bits in one frame interval, future frames are not encoded, but skipped. Here, M can correspond to the buffer capacity, where $M=R/F$ is a suitable choice.[10] The target number of bits per frame is specified as $B_{TR}=R/F-\Delta$, where the correction factor

$$\Delta=\begin{cases} W/F, & W>0.1M \\ W-0.1M, & \text{otherwise} \end{cases}$$

provides feedback to maintain buffer fullness at $W=0.1\,M$, while the average number of bits per frame over a number of frames is approximately R/F. Note that if the target buffer fullness of $0.1\,M$ is met after a frame is encoded, and R/F bits are emptied out, then the correction factor becomes $\Delta=0$ for the next frame.

The results for the TMN5 encoder incorporating the two variations (***New*** and ***New***$^{\kappa}$) of the proposed macroblock layer rate control method are presented and compared with the results for the TMN5 encoder incorporating the macroblock layer rate control methods of TMN5[17] and TMN8.[8–10]

For the implementation of the proposed method $l_{\max}=100$, $\delta=4$ were used in Eq. (2) to get a good tradeoff between preserving the variance information and keeping the classification complexity low. A universal lookup table of estimates was created offline by the application of Eq. (5) with the number of bits expended for the macroblocks of the first ten frames of the training sequences News, Mother and Daughter, Hall Objects, and Susie, sampled at frame rates 30, 15, 10, and 7.5 Hz, and quantized uniformly with the same quantization parameter, assuming all values in the range 1, . . . ,31. Before the encoding of a test sequence, the number of occurrences for each combination of class c and quantization parameter q was initialized to a small value [i.e., $P_0(c,q)=0.1$], if the corresponding table entry had been populated with training data. Table entries were (adaptively) updated after the encoding of each frame of the test sequences by the application of Eq. (5), with the number of bits expended for the macroblocks of that frame. Choice of $P_{\max}=512$ in Eq. (6) enabled sufficiently fast adaptivity of the table of estimates to the incoming data.

Tables 1 and 2 show the rms and the maximum, respectively, of the deviation of the actual number of expended bits from the target number of bits for the frames of the test sequences coded with all of the rate control methods. The accuracy of the proposed method is seen to be superior to the accuracy of the TMN5 rate control method for each experiment, and slightly better than the accuracy of the TMN8 rate control method on the average. The variation ***New***$^{\kappa}$ exhibits a distinct performance advantage over the variation ***New*** for test cases of Coastguard coded at 10 and 30 frames/sec, and Carphone, Container, Foreman, and Silent coded at 10 frames/sec. In all these cases, the rate contribution of motion vector estimates is large.

The number of bits encoded with the variation ***New*** and the TMN8 method is seen to be nearly constant for each frame of the Foreman sequence coded at 48 Kbits/sec and 128 Kbits/sec, as shown in Fig. 3. As a consequence of transmitting the desired number of bits with these methods, the resulting buffer fullness could be reduced to the target fullness of 10% of the frame skip threshold and/or maintained at this low level with these methods, as shown in Fig. 4. The large variations, seen in Fig. 4, in the buffer fullness level for the rate control method of TMN5 sometimes resulted in additional frame skips in some of these experiments when the frame skip threshold was exceeded. The method of TMN8 and the variation ***New*** of the proposed method did not require any additional frame skips. For scenes where the human eye can track uniform motion,

Table 2 Maximum deviation of the number of actual bits from the target number of bits for a frame.

	Kbits/sec, Hz	Akiyo	Carphone	Claire	Coast	Container	Foreman	Silent
TMN5	48, 10	1728.000	1725.600	1578.399	792.0000	1160.000	1866.399	1477.600
TMN8	48, 10	256.0000	571.2001	122.3999	75.20019	1296.000	768.0000	280.0000
New	48, 10	265.6000	188.7998	120.0000	376.0000	280.0000	759.2001	142.3999
New^{K}	48, 10	392.0000	72.00000	111.2001	75.20019	137.6000	728.0000	136.0000
New (No update)	48,10	2340.799	1588.799	1739.200	2492.799	1536.000	2368.799	2008.000
TMN5	128, 30	3195.333	2316.666	2840.399	2105.399	2770.466	1572.666	3153.533
TMN8	128, 30	420.3999	243.6000	417.6665	91.33349	1791.333	344.9331	353.8666
New	128, 30	381.3334	116.6665	260.2666	330.6665	281.9331	230.6665	269.7998
New^{K}	128, 30	361.5332	92.60009	268.6665	102.3999	324.6000	188.0668	312.6665
New (No update)	128,30	2897.133	1396.333	1934.133	2641.666	1537.333	1371.466	3106.133

frame skipping is easily perceptible. Frame skipping may not be tolerable in sign language or lip-reading video applications.

As seen in Fig. 4, the experiment of the coding of Foreman at 48 Kbits/sec and 10 frames/sec reveals that the proposed method is slightly more robust than the method of TMN8 in high-stress conditions, such as the rapid camera movement in the middle of this sequence.

One can also put the rate control accuracy of the proposed method into better perspective by referring to Ribas-

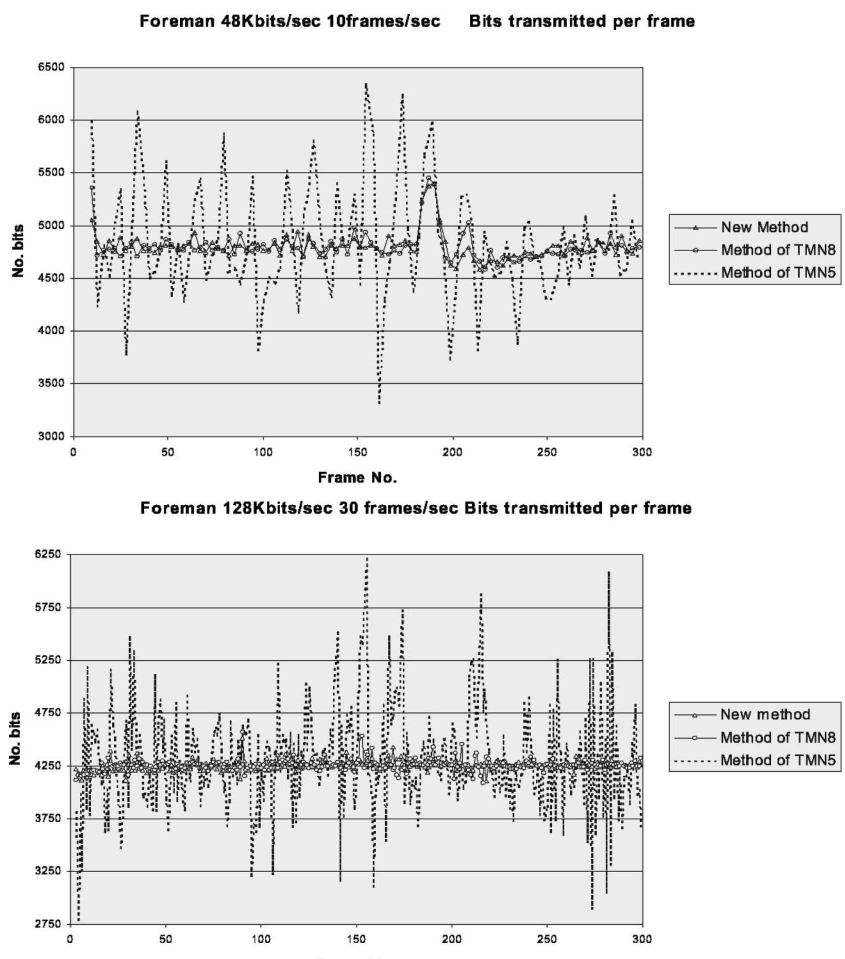

Fig. 3 Plot of the number of bits transmitted (encoded) for each frame of the Foreman sequence. Two frames were skipped in the 48 Kbits/sec, 10 Hz case and four frames were skipped in the 128 Kbits/sec, 30 Hz case with the TMN5 rate control method. No frame skips occurred for either case with the proposed and TMN8 rate control methods.

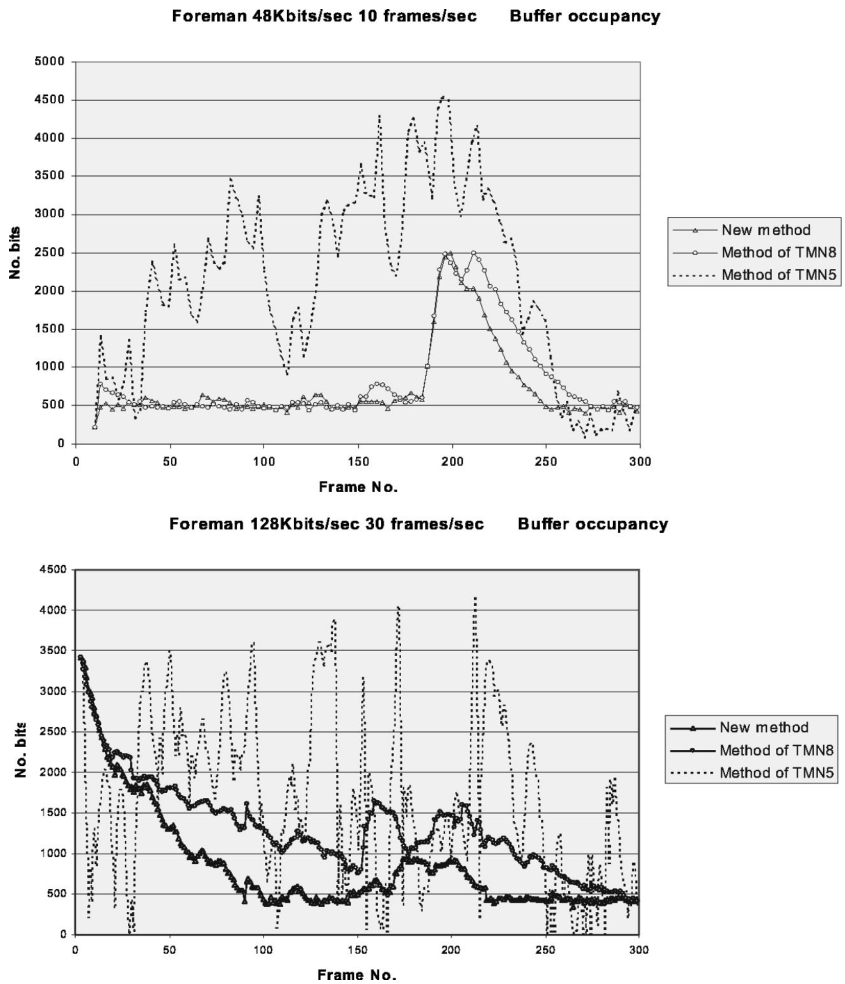

Fig. 4 Number of bits remaining in the buffer after the transmission of each frame.

Corbera and Lei,[9] where the TMN8 macroblock layer rate control[8–10] in the MPEG-4 codec is reported to yield a bit rate closer to the target than that of the VM7[6] macroblock layer rate control with far fewer frame skips as well.

Adaptive updating of the quantization parameter values after the encoding of each macroblock, as discussed in the last paragraph of Sec. 3, is the feature of the proposed method that extends the work of Puri and Aravind[15] for tight buffer regulation. The last rows of Tables 1 and 2 demonstrate the importance of this feature by showing that when this feature is disabled, the resulting deviations from the targets are comparable to the deviations for the TMN5 method at high frame rates and exceeds the deviations for the TMN5 method at low frame rates.

Table 3 shows that on the average, the two variations of the proposed method outperform the method of TMN5 in terms of average reconstructed P frame (luminance) PSNR, mainly due to the better utilization of the buffer. The TMN5

Table 3 Average P frame reconstruction PSNR for luminance components obtained with rate control methods.

	Kbits/sec, Hz	Akiyo	Carphone	Claire	Coast	Container	Foreman	Silent
TMN5	48, 10	40.32836	32.71587	41.19091	28.89683	35.11866	30.06927	34.23884
TMN8	48, 10	40.70438	32.71185	41.39898	28.89479	35.44360	30.00732	34.16804
New	48, 10	40.75367	32.77463	41.45683	28.94061	35.58762	30.06463	34.22925
New$^{\kappa}$	48, 10	40.74153	32.78206	41.43561	28.94183	35.51567	30.08371	34.22804
TMN5	128, 30	42.70044	33.66689	43.29551	30.35431	36.95077	31.97853	36.71260
TMN8	128, 30	42.81043	33.59821	43.29875	30.29700	37.02414	31.80528	36.64921
New	128, 30	43.12215	33.66235	43.46402	30.36023	37.28148	31.87606	36.79819
New$^{\kappa}$	128, 30	43.21963	33.65882	43.44765	30.36161	37.29986	31.87794	36.78997

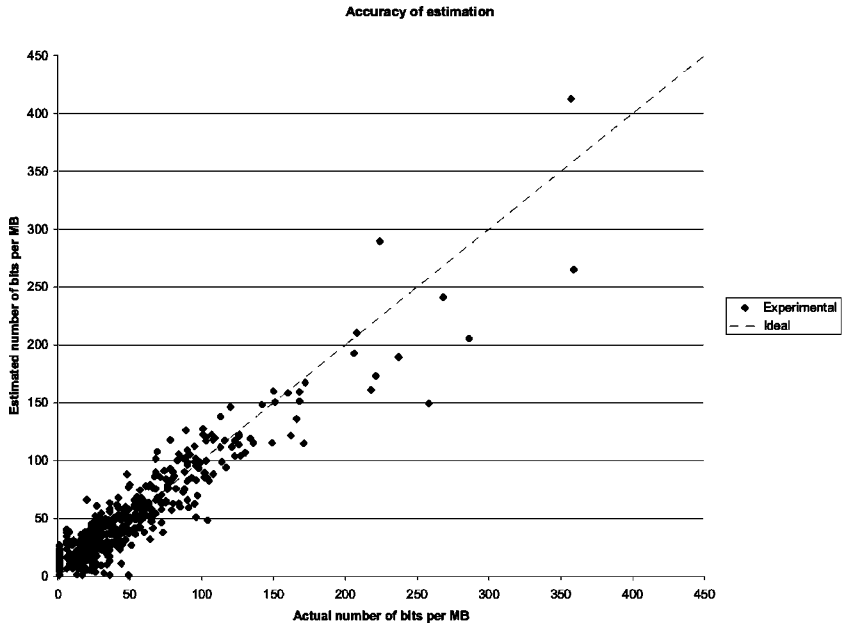

Fig. 5 The scatter plot of the estimated number of coding bits versus the actual number of coding bits for five frames of the Foreman sequence (48 Kbits/sec, 10Hz) confirms that the estimation process has a near-zero bias and a small estimation error variance (data points are centered and clustered around a line of unity slope).

rate control method frequently yields an average coding rate that is far lower than the target rate, whereas the proposed and the TMN8 rate control methods can achieve the target rate almost exactly. For example, for the Container sequence at 30 Hz, actual average rates with the rate control methods of TMN5, TMN8, and ***New*** are 122.90, 127.85, and 128.06 Kbps, respectively, for a target rate of 128.00 Kbps. In the only test case of Foreman (128 Kbps, 30 Hz), where the average reconstructed P frame PSNR value for the TMN5 rate control method has exceeded that of the proposed method, the TMN5 rate control method skipped four frames, and the average number of bits (4306 bits) that were coded per frame with the TMN5 rate control was more than the targeted average (128000/30 $=4267$ bits).

On the other hand, the fact that the average reconstructed P frame PSNR is consistently better with the two variations of the proposed method than with the method of TMN8 suggests that the conceptually simple strategy of near-uniform quantization parameter assignment could rival the rate-distortion optimized assignment[8–10] strategy, utilizing approximate rate versus quantization parameter and distortion versus quantization parameter relations.

One can put the objective video quality advantage of the proposed method into better perspective by working with bit rate savings rather than PSNR gains. For example, for Akiyo, coded by TMN5 and TMN8 at 48 Kbits/sec and 10 Hz, the PSNR gains in Table 3 translate to bit rate savings of 8.3 and 0.8%, respectively. For Coastguard, coded by TMN5 and TMN8 at 48 Kbits/sec and 10 Hz, the bit rate savings are 1.1 and 1.5%, respectively.

The method of nonlinear MMSE bit count estimation, which forms the basis of the proposed rate control method, yields a near-zero bias as well as a small error variance at all rates, as shown in Fig. 5. In this figure, the estimated number of coding bits is plotted against the actual number of coding bits for each macroblock of five equispaced coded frames taken from the Foreman sequence. The estimation method can also accurately predict noncoded macroblocks, where the number of estimated coding bits is usually close to 1.

6 Complexity Issues

Insights into the complexity of the proposed method can be gained by analyzing the method in several steps. As with most other macroblock layer rate control methods, which exploit macroblock data content,[5,7–10,15] a macroblock statistic needs to be computed for all the macroblocks in the first step of bit count estimation. This either implies the motion estimation to be completed for all the macroblocks before the DCT, quantization, and coding, or the motion estimation to be performed twice, once for the purpose of rate control and once for the actual coding of the macroblock. Hence memory or speed requirements increase over the simple TMN5 rate control method.[17] However, we note that this first step is also performed in the TMN8 rate control method.

In the second and third steps of bit count estimation, the scalar quantization of the macroblock statistic is implemented by comparison operations, and the mapping of the class to a bit count estimate is implemented by a table lookup operation for each quantization parameter value. The total complexity of these stages is not more than that of the evaluation of the parametric rate versus quantization parameter equations in other methods,[4–11] which require multiplication and division operations.

Finally, for quantization parameter assignment, as with some of the other methods,[4–7,11] a search for the optimum (combination) of quantization parameter values for the macroblocks, yielding a total bit count estimate closest to

Table 4 Average execution time for the processing of P frames of the Foreman sequence coded at 48 Kbits/sec, 10 Hz with the rate control methods (figures obtained by averaging six runs).

Method	TMN5	TMN8	***New***	***New***$^{\kappa}$
Time(sec)	0.0378	0.0410	0.0423	0.0425

the target number of bits for the frame, must be conducted. An exhaustive search requires the bit count estimates to be made for all values of the quantization parameter that can be assigned to each macroblock. Let N_q be the number of quantization parameter values that can be assigned to each one of Z macroblocks of a frame in which Z_1 macroblocks have already been coded. Then the upper bound on the number of different combinations of quantization parameter assignments that will be considered is $N_q^{Z-Z_1}$. The near-uniform quantization parameter assignment process of Sec. 3 lowers this upper bound down to $N_q(Z-Z_1)$ combinations by permitting up to one step size transition of the quantization parameter per frame. The constrained search, which needs to be repeated after each encoded macroblock, is nevertheless computationally less attractive than the method of TMN8,[8–10] which directly yields the quantization parameter given the standard deviations of the macroblocks.

In short, the computational complexity of the proposed method is mainly due to the number of bit count estimations performed and the search for the optimum combination of quantization parameters rather than the complexity of each bit count estimation.

In Table 4, we list the average encoding time of the P frames of the 48 Kbits/sec, 10 frames/sec coded Foreman sequence with all four rate control methods. The increase in computational complexity with the proposed method is less than 12.5% that of the computational complexity of the method of TMN5, and less than 4% that of the computational complexity of the method of TMN8.

7 Conclusion

A low-delay H. 263 video rate control method, based on the classification of the macroblock data content and the subsequent mapping of the macroblock class to a nonlinear estimate of the number of bits to be expended for the macroblock, is proposed. The design and the periodic update of the nonlinear bit count estimate for a particular class and quantization parameter combination is based on the measured number of bits expended for previously encoded macroblocks of the same class coded with the same quantization parameter. Given the bit count estimates for each combination of macroblock class and quantization parameter, the quantization parameter values are assigned to macroblocks by permitting up to one step transition within a frame. This conceptually simple strategy of near-uniform assignment of quantization parameters to macroblocks yields a high objective spatial quality, and the update of the quantization parameters after each encoded macroblock permits the target number of bits for the frame to be accurately met.

Even though the macroblock classification-based bit count estimation in the proposed rate control method is similar to that of Puri and Aravind,[15] the proposed method can achieve high objective spatial quality and improved rate control accuracy for tighter buffer regulation due to the following key features. 1. Coding mode as well as macroblock data content is used in macroblock classification. 2. The sum of all quantization bits and some overhead bits is estimated and combined with the number of remaining overhead bits, which are directly measured. This process generally yields more accuracy than the direct estimation of the number of all coding bits of a macroblock. 3. The quantization parameters are adjusted to meet the remaining bit budget for the frame after the encoding of each macroblock.

The performance of the proposed rate control method has been demonstrated to be better than the methods of TMN5[17] and TMN8[8–10] in terms of average P frame reconstruction PSNR and bit rate control accuracy. This comes at the expense of an affordable computational complexity increase over the methods of TMN5 and TMN8 due to the quantization parameter assignments.

References

1. K. T. Ng, S. C. Chan, and T. S. Ng, ''Buffer control algorithm for low bit-rate video compression,'' Proc. ICIP pp. 685–688 (Sep. 1996).
2. G. Schuster and A. Katsaggelos, ''Fast and efficient mode and quantizer selection in the rate-distortion sense for H.263,'' *Proc. SPIE* **2727**, 784–795 (Mar. 1996).
3. D. Mukherjee and S. K. Mitra, ''Combined mode selection and macroblock quantization step adaptation for the H.263 video encoder,'' *Proc. ICIP* **2**, 37–40 (Oct. 1997).
4. K. Yang, A. Jacquin, and N. Jayant, ''A normalized rate-distortion model for H.263-compatible codecs and its application to quantizer selection,'' *Proc. ICIP* **2**, 41–44 (Oct. 1997).
5. E. Viscito and C. Gonzales, ''A video compression algorithm with adaptive bit allocation and quantization,'' *Proc. SPIE* **1605**, 58–71 (Nov. 1991).
6. T. Chiang and Y.-Q. Zhang, ''A new rate control scheme using quadratic rate distortion model,'' *IEEE Trans. Circuits Syst. Video Technol.* **7**(1), 246–250 (Feb. 1997).
7. H.-J. Lee, T. Chiang, and Y.-Q. Zhang, ''Scalable rate control for MPEG-4 video,'' *IEEE Trans. Circuits Syst. Video Technol.* **10**(6), 878–894 (Sep. 2000).
8. J. Ribas-Corbera and S. Lei, ''Rate control for low delay video communications,'' ITU-T Study Group 16 Doc. Q15-A-20 (June 1997).
9. J. Ribas-Corbera and S. Lei, ''Rate control in DCT video coding for low-delay communications,'' *IEEE Trans. Circuits Syst. Video Technol.* **9**(1), 172–185 (Feb. 1999).
10. ITU-T Study Group 16, *Video Codec Test Model, Near Term, Version 8 (TMN8) Release 0*, ITU-T Study Group 16 Doc. Q15-A-59 (June 1997).
11. W. Ding and B. Liu, ''Rate control of MPEG video coding and recording by rate quantization modelling,'' *IEEE Trans. Circuits Syst. Video Technol.* **6**(1), 12–20 (Feb. 1996).
12. Z. He, Y. K. Kim, and S. K. Mitra, ''ρ-domain source modelling and rate control for video coding and transmission,'' *Proc. ICASSP* **3**, 1773–1776 (May 2001).
13. Z. He and S. K. Mitra, ''Optimum bit allocation and accurate rate control for video coding via ρ-domain source modelling,'' *IEEE Trans. Circuits Syst. Video Technol.* **12**(10), 840–849 (Oct. 2002).
14. S. Bilato, G. Calvagno, G. A. Mian, and R. Rinaldo, ''Accurate bitrate and quality control for the MPEG video coder,'' *Proc. ICIP* **3**, 571–574 (Oct. 1997).
15. A. Puri and R. Aravind, ''Motion compensated video coding with adaptive perceptual quantization,'' *IEEE Trans. Circuits Syst. Video Technol.* **1**(4), 351–361 (Dec. 1991).
16. ISO-IEC/JTC1/SC29/WG11, *Test Model 5*, JTC1/SC29/ WG11 Coding of Moving Pictures and Associated Audio (1994).
17. ITU-T SG 15, Working Party 15/1, *Video Codec Test Model, TMN5*, Telenor Research (Jan. 1995).
18. K Oehler and J. L. H. Web, ''Macroblock quantizer selection for H.263 video coding,'' *Proc. ICIP* **1**, 365–369 (Oct. 1997).

19. A. Nicoulin, M. Matavelli, W. Li, A. Basso, A. C. Popat, and M. Kunt, ''Image sequence coding using motion-compensated subband decomposition,'' in *Motion Analysis and Image Sequence Processing*, Sezan and Lagendijk, Eds., pp. 225–256, Kluwer Academic Publishers, Norwell, MA (1993).
20. S. Daly, K. Matthews, and J. Ribas-Corbera, ''Face-based visually-optimized image sequence coding,'' *Proc. ICIP* **3**, 443–447 (1997).
21. Y. Shoham and A. Gersho, ''Efficient bit allocation for an arbitrary set of quantizers,'' *IEEE Trans. Acoust., Speech, Signal Process.* **36**, 1445–1453 (Sep. 1988).

Ulug Bayazit received his BS degree in electrical engineering from Bogazici University, Istanbul, Turkey, in 1991. He received MS degrees in electrical engineering and mathematics in 1993 and 1995, respectively, and his PhD degree in electrical engineering in 1996 from Rensselaer Polytechnic Institute, Troy, New York. Between 1996 and 2000 he was employed as a research scientist and a systems architect for Toshiba America. He joined the electronics engineering department of Isik University in Istanbul, Turkey, in 2000, where he is currently an assistant professor. His research interests are in data compression, and image, video, and multimedia coding and processing.